UNIVERSITI TEKNOLOGI MARA

LINEAR-PSO WITH BINARY SEARCH ALGORITHM FOR DNA MOTIF DISCOVERY

HAZARUDDIN BIN HARUN

Thesis submitted in fulfillment
of the requirements for the degree of
Doctor of Philosophy

Faculty of Computer and Mathematical Sciences

January 2015

CONFIRMATION BY PANEL OF EXAMINERS

I certify that a panel of examiners has met on 27th June 2014 to conduct the final examination of Hazaruddin Bin Harun on his Doctor of Philosophy thesis entitled "Linear-PSO with binary search algorithm for DNA motif discovery" in accordance with Universiti Teknologi MARA Act 1976 (Akta 173). The Panel of Examiners recommends that the student be awarded the relevant degree. The panel of Examiners was as follows:

Zainab Abu Bakar, PhD
Professor
Faculty of Computer & Mathematical Sciences
Universiti Teknologi MARA
(Chairman)

Noor Elaiza Abd. Khalid, PhD
Senior Lecturer
Faculty of Computer & Mathematical Sciences
Universiti Teknologi MARA
(Internal Examiner)

Tengku Mohd Tengku Sembok , PhD
Professor
Faculty of Defence Science and Technology
Universiti Pertahanan Nasional Malaysia
(External Examiner)

Graham Kendall, PhD
Professor
School of Computer Science
The University of Nottingham Malaysia Campus
(External Examiner)

SITI HALIJAH SHARIFF, PhD
Associate Professor
Dean
Institute of Graduates Studies
Universiti Teknologi MARA
Date: 7 January 2015

AUTHOR'S DECLARATION

I declare that the work in this thesis was carried out in accordance with the regulations of Universiti Teknologi MARA. It is original and is the result of my own work, unless otherwise indicated or acknowledged as referenced work. This thesis has not been submitted to any other academic institution or non-academic institution for any degree or qualification.

I, hereby, acknowledge that I have been supplied with Academic Rules and Regulations for Post Graduate, Universiti Teknologi MARA, regulating the conduct of my study and research.

Name of Student	:	Hazaruddin Bin Harun
Student I.D. No.	:	2008531847
Programme	:	Doctor of Philosophy (CS990)
Faculty	:	Faculty of Computer and Mathematical Sciences
Thesis Title	:	Linear-PSO with Binary Search Algorithm for DNA Motif Discovery
Signature of Student	:	.....
Date	:	January 2015

ABSTRACT

Motif Discovery (MD) is the process of identifying meaningful patterns in DNA, RNA, or protein sequences. In the field of bioinformatics, a pattern is also known as a motif. Numerous algorithms had been developed for MD, but most of these were not designed to discover species specific motifs used in identifying a specifically selected species where the exact location of these motifs also needs to be identified. Evaluation of these algorithms showed that the results are unsatisfactory due to the lower validity and accuracy of these algorithms. At present, DNA sequencing analysis is the most utilised technique for species identification where patterns of DNA sequences are determined by comparing the sequence to comprehensive databases. However, several false and gap sequences had been identified to be present in these databases which lead to false identification. Therefore, this study addresses these problems by introducing a hybrid algorithm for MD. In this study, the MD is a process to discover all possible motifs that existed in DNA sequences whereas Motif Identification (MI) is a process to identify the correct motif that can represent a selected species. Particle Swarm Optimisation (PSO) was selected as the base algorithm that needs improvement and integration with other techniques. The Linear-PSO algorithm was the first version of improvement. However due to the longer time required for complete execution of this algorithm, the Binary Search technique was integrated and a new version of the algorithm was developed, namely the Linear-PSO with Binary Search (LPBS) algorithm. A total of 11 experiments were conducted in this research, where the aim of the first four experiments was algorithm improvement; the next four experiments were for identifying suitable input data, while the final three experiments were for algorithm validation. Several DNA sequences from different species were collected from the GenBank and TRansCompel databases and used as input for the algorithm. The collected DNA sequences were from the Mitochondrial Cytochrome C Oxidase Subunit I (COX1) gene. Due to the limitation of available data, only four species were collected for Motif Discovery, namely pig, cow, yak, and chicken. Another five species were used for Motif Identification, which were human, sheep, dog, frog, and rat. The algorithm was run on an Intel(R) Core(TM) Duo CPU 1.73 GHz notebook with 3 GB RAM. The results showed that the LPBS algorithm was able to discover possible correct motifs that can represent a species with higher validity and accuracy as compared to previous algorithms. The motifs discovered were consistent for each execution with higher calculated fitness values.

ACKNOWLEDGEMENT

In the name of Allah, the Most Gracious and the Most Merciful; Shalawat and Salam to Rasulullah Muhammad (s.a.w). Alhamdulillah, all praises to Allah, who instilled in me the courage and provided guidance for me to complete this research.

I would like to thank all those who, directly and indirectly, made this exploration of thesis and production of this dissertation a reality, which is a unique and unforgettable experience for me.

First and foremost, I would like to express my sincerest gratitude to my supervisor Associate Professor Dr. Sharifah Lailee Syed Abdullah for her supervision and support in the completion of my doctoral dissertation. Also my deepest thanks to my co-supervisor Associate Professor Dr. Naimah Mohd Hussin, for her help and knowledge sharing related to this topic of research.

My sincerest thanks and acknowledgement to UiTM and MOHE for providing the necessary financial support that funded parts of the research discussed in this dissertation.

I would also like to thank my parents for their support and encouragement with their best wishes and *doa*.

My deepest appreciation goes to my wife, Hamirul'Aini Hambali and my children for their never-ending patience and understanding.

Finally, I would like to offer my special thanks to those who are not mentioned in this text, but have contributed to the success of this doctoral journey.